Gruca *et al. BMC Bioinformatics* (2021) 22:182
https://doi.org/10.1186/s12859-021-04017-7

BMC Bioinformatics

RESEARCH ARTICLE

Open Access

# Common low complexity regions for SARS-CoV-2 and human proteomes as potential multidirectional risk factor in vaccine development

Aleksandra Gruca[1], Joanna Ziemska-Legiecka[2], Patryk Jarnot[1], Elzbieta Sarnowska[3], Tomasz J. Sarnowski[2] and Marcin Grynberg[2*]

*Correspondence: greenb@ibb.waw.pl
[2] Institute of Biochemistry and Biophysics, Polish Academy of Sciences, Warsaw, Poland
Full list of author information is available at the end of the article

## Abstract

**Background:** The rapid spread of the COVID-19 demands immediate response from the scientific communities. Appropriate countermeasures mean thoughtful and educated choice of viral targets (epitopes). There are several articles that discuss such choices in the SARS-CoV-2 proteome, other focus on phylogenetic traits and history of the Coronaviridae genome/proteome. However none consider viral protein low complexity regions (LCRs). Recently we created the first methods that are able to compare such fragments.

**Results:** We show that five low complexity regions (LCRs) in three proteins (nsp3, S and N) encoded by the SARS-CoV-2 genome are highly similar to regions from human proteome. As many as 21 predicted T-cell epitopes and 27 predicted B-cell epitopes overlap with the five SARS-CoV-2 LCRs similar to human proteins. Interestingly, replication proteins encoded in the central part of viral RNA are devoid of LCRs.

**Conclusions:** Similarity of SARS-CoV-2 LCRs to human proteins may have implications on the ability of the virus to counteract immune defenses. The vaccine targeted LCRs may potentially be ineffective or alternatively lead to autoimmune diseases development. These findings are crucial to the process of selection of new epitopes for drugs or vaccines which should omit such regions.

**Keywords:** COVID-19, SARS-CoV-2, Human, Coronavirus, Epitope, Low complexity region, Sequence conservation, Spike glycoprotein (s), nsp3, Nucleocapsid protein (n)

## Background

At the very end of 2019 the Chinese Center for Disease Control (China CDC) reported several severe pneumonia cases of unknown etiology in the city of Wuhan. The causative agent of the disease was a previously unknown *Betacoronavirus* named SARS-CoV-2. The virus quickly spread all over the globe (https://www.who.int/emergencies/diseases/novel-coronavirus-2019; https://www.worldometers.info/coronavirus/) and as of today

(June 2020) the number of infections and the number of deaths were globally still on the rise.

Coronaviruses are widespread in vertebrates and cause a plethora of respiratory, enteric, hepatic, and neurologic issues. Some of the animal coronaviruses exhibited ability to transmit to human e.g. the severe acute respiratory syndrome coronavirus (SARS-CoV) in 2003 and Middle East respiratory syndrome coronavirus (MERS-CoV) in 2012 had caused human epidemics [1, 2]. SARS-CoVs enters cells via the angiotensin-converting enzyme 2 (ACE2) receptor [3, 4]. The SARS-CoV-2 first infects airways and binds to ACE2 on alveolar epithelial cells. Both viruses are potent inducers of inflammatory cytokines [5]. The virus activates immune cells and induces the secretion of inflammatory cytokines and chemokines into pulmonary vascular endothelial cells. In severe cases of COVID-19 the patient develops a "cytokine storm" [6–9]. Since most infected individuals are apparently asymptomatic it is hard to assess the prevalence of SARS-CoV-2 in global or even local populations. Lack of appropriate testing quantities also plays a role. To date there are no fully effective drugs or vaccines against SARS-CoV-2 [6, 10–14].

Several recent articles have suggested that SARS-CoV-2 proteins and some protein domains are important to the viral lifecycle, several of which are conserved in the Coronaviridae family and which are possible targets as epitopes [15–22]. This is of importance since the choice of proper target/epitopes is crucial for drug or vaccine design.

The quest for optimal epitope targets is difficult and focuses on both experimental and bioinformatic means. This may directly involve laboratory experiments or databases like the Immune Epitope Database and Analysis Resource (IEDB)[16, 22, 23]. The second bioinformatic approach uses the search for similarities with other viruses in order to identify conserved regions of the viral genome [19, 24, 25].

Although both methods focus on the viral RNA/proteome sequence, these approaches clearly treat this sequence as a whole. However, one can clearly distinguish between protein high complexity regions (HCRs) encompassing most of the genome and low complexity regions (LCRs). LCRs are often described as 'unstructured' or are simply not annotated. Our recent experiments however show that the search for similar sequences in LCRs is almost impossible using standard methods, like BLAST or HHblits [26, 27]. This is why we created three algorithms that are able to compare low complexity protein sequences, GBSC, MotifLCR and LCR-BLAST [28–32].

The current situation due to COVID-19 is critical, and this work aims to compare the SARS-CoV-2 LCRs with the human proteome. In this work we show that numerous fragments of the viral proteome are very similar to the human ones. It has been recently shown that the furin cleavage site of Spike SARS-CoV-2 protein shares similarity with the human epithelial sodium channel [33]. Our findings suggest that identified fragments of spike, nsp3 and nucleocapsid proteins should not be considered as epitopes neither for vaccine nor drug design.

Moreover, our hypothesis is supported by the malaria molecular evolution and vaccine design study clearly indicating that LCRs may play a role in immune escape mechanism [34]. The attempt to develop about 100 anti-malaria vaccine candidates indicated their limited protective effect. The global analysis of antigens led to the conclusion that LCRs present in proteins containing glutamate-rich and/or repetitive motifs carried the most immunogenic epitopes. On the other hand the antibodies recognizing these epitopes

appeared to be ineffective in an in vitro study [35]. Moreover, the exhaustive study showed that LCRs may drive the immune response away from important functional domains in parasite proteins [35]. In this context it seems to be very important to avoid the presence of LCRs in vaccine epitopes due to 1) low effectiveness of antibodies recognizing this region, even if LCRs are highly immunogenic and 2) the presence of LCRs in some human proteins. Additionally, the deep study revealed that molecular mimicry may serve as an attractive explanation of autoimmune side effects after pathogen infections [36]. Moreover, it was already reported in COVID-19 that some patients developed the self-reacting antibody like e.g. anti-nuclear antibodies, anti-phospholipids antibody, anti-INF antibodies and anti-MDA5 antibodies [37]. Further, new clinical reports indicate the presence of autoantibodies which can reach the brain. In patients with severe COVID-19 the blood–brain barrier dysfunction was detected as well as the neuronal damage was found and increased levels of self-reacting antibodies in cerebrospinal fluid. These antibodies mostly recognized the epitopes in the brain. Finally, the fraction of patient-derived virus neutralizing monoclonal antibodies can recognize the targets in mammalian cells [38]. So far there is no evidence that LCRs may cause the self-reactivity of the immune system in other infections but concerning the high number of autoimmune side effects after SARS-CoV-2 infection we can not exclude such possibility.

## Results

Our aim was to find the LCRs in the SARS-CoV-2 genome that are similar to fragments of human proteins and to identify if any of those overlap with other epitopes in an attempt to eliminate epitope hits that are too similar to the human proteome fragments.

### Protein similarity between SARS-CoV-2 and human LCRs

To achieve this goal we used our three methods: GBSC, MotifLCR and LCR-BLAST. GBSC takes as an input whole protein sequences, however the input for LCR-BLAST and MotifLCR is expected to consist of LCRs. Therefore to identify LCRs in the SARS-CoV-2 and in the human proteomes we used the SEG tool with default parameters. The detailed description of the data sources and the methods used to identify and analyze low complexity regions in SARS-CoV-2 and human proteome is provided in Supplemental Material 9 “Materials and Methods”. There are 23 LCRs in SARS-CoV-2 proteome that were found in the following proteins: nsp2 (636 aa, one LCR found), nsp3 (1945 aa, six LCRs found) nsp4 (500 aa, one LCR found), nsp6 (290 aa, one LCR found), nsp7 (83 aa, two LCRs found), nsp8 (198 aa, two LCRs found), S protein (1273 aa, two LCRs found), E protein (75 aa, one LCR found), orf7a (121 aa, one LCR found), orf7b (43 aa, one LCR found), N protein (419 aa, four LCRs found) and orf14 (73 aa, one LCR found) (Fig. 1). It is worth noting that most LCRs are located either in pp1a or in the C-terminal proteins (from spike to nucleocapsid protein). The middle section, from nsp9 to nsp16 is completely devoid of such sequences. In the next step we identified which of the SARS-CoV-2 LCRs are similar to human LCRs (Fig. 1). Similar fragments are present in nsp3, spike glycoprotein and in the nucleocapsid protein. The list of these regions is presented in Table 1. We also provide a list of similar protein fragments from the human proteome obtained with three different methods (see Additional files 1, 2, 3: S1-3 Tables).


**Fig. 1** SARS-CoV-2 proteins shown according to their encoding position in the genome. Low Complexity Regions (LCRs) are marked with triangles. LCRs that are similar to human proteins are highlighted in red. The original figure for this modification was kindly provided by ViralZone, SIB Swiss Institute of Bioinformatics (https://viralzone.expasy.org/8996)

**Table 1** List of SARS-CoV-2 low complexity regions that are similar to human proteins

| LCR sequence | potein | LCR start | LCR end |
|---|---|---|---|
| PPDEDEEEGDCEEEEFE | nsp3 | 108 | 124 |
| QPEEEQEEDWLDDDSQ | nsp3 | 152 | 167 |
| MLCCMTSCCSCLKGCCSCGSCC | S protein | 1233 | 1254 |
| GSRGGSQASSRSSSRSRNSSRNSTPGSSRGTSP | N protein | 175 | 207 |
| KTFPPTEPKKDKKKKADE | N protein | 361 | 378 |

Figure 1 SARS-CoV-2 proteins shown according to their encoding position in the genome. Low Complexity Regions (LCRs) are marked with triangles. LCRs that are similar to human proteins are highlighted in red. The original figure for this modification was kindly provided by ViralZone, SIB Swiss Institute of Bioinformatics (https://viralzone.expasy.org/8996)

Similarity of nsp3 is most significant to the myelin transcription factor 1-like protein (Myt1l) (Table 1, Additional files 1, 2, 3: S1-S3 Tables). Myt1l was shown to be expressed in neural tissues in the developing mouse embryo [39]. Myt1l is supposed to limit non-neuronal genes expression, take part in neurogenesis and functional maintenance of mature neurons [40]. The glutamic acid-rich fragment is located close to the activation domain, however it was shown to be dispensable in this process [41].

The spike glycoprotein fragment MLCCMTSCCSCLKGCCSCGSCC has significant similarity to LCRs of ultrahigh sulfur keratin-associated proteins present both in hair cortex and cuticle (KRTAP 4.3, KRTAP 5.4 and KRTAP 5.9) [42] (Table 1, Additional files 2, 3: S2 and S3 Tables). KRTAPs are parts of the intermediate filaments of the hair shaft.

Nucleocapsid protein (N) has 2 LCRs that are similar to human LCRs (Table1, Additional files 1, 2, 3: S1-S3 Tables) the most interesting comparable fragment is the zinc finger Ran-binding domain-containing protein 2 (RANB2), which is a part of the supraspliceosome where it is responsible for alternative splicing [43, 44]. GBSC identifies the high similarity of the viral LCR to a LCR of the solute carrier family 12 which is an electroneutral potassium-chloride co-transporter which can be mutated in some severe peripheral neuropathies [45, 46]. The C-terminal LCR is similar to a LCR from [F-actin]-monooxygenase MICAL3, actin-regulatory redox enzyme that directly binds and disassembles actin filaments (F-actin) [47]. This protein is also responsible for exocytic vesicles tethering and fusion, and cytokinesis [48–50]. The region of interest is probably involved in binding some of a multitude of binding partners of MICAL3 [49] (https://www.ebi.ac.uk/intact/interactors/id:Q7RTP6*).

Lists of human hits of LCRs similar to viral fragments were annotated with Gene Ontology (GO) terms [51] in order to find common functional features that were over-represented among proteins composing the clusters. Here we focus on the results for GO annotations from the Biological Processes namespace since these functions may be crucial to understanding possible viral interventions into the cellular machinery. Complete lists of enriched GO terms are available in Additional files 4, 5, 6, 7, 8: Tables S4-S8. The best matches for the first LCR in nsp3 are human proteins involved in actin processing (Additional files 4: Table S4). The best matches for the adjacent LCR in nsp3 are related to signal transduction (Table S5). The best hits for the spike protein LCR fragment are related to keratin (Additional files 6: Table S6). The human proteins similar to the central nucleocapsid protein's LCR show discrepancies between sets of results. The output of GBSC clearly points to salinity response/salt stress responses (Additional files 7: Table S7). Results from LCR-BLAST and MotifLCR are actin-centred (Additional files 7: Table S7). In the case of the C-terminal nucleocapsid protein LCR, the most abundant human representatives are exocytosis and oxidation–reduction processes (Additional files 8: Table S8).

#### Motif similarities of SARS-CoV-2 and human LCRs

We also tested similarities of viral LCR fragments to known domains and motifs using the UniProt, PROSITE, CDD, InterPro and ELM databases [52–56]. Most of the matches to known domains and motifs of the SARS-CoV-2 LCRs are to previously annotated regions, *i.e.* compositionally biased regions, rich in a particular amino acid or polyX regions. Only in two cases are there hits to specific domains.

The first similarity between SARS-CoV-2 LCR and a known motif is between the surface glycoprotein LCR (MLCCMTSCCSCLKGCCSCGSCC) and the keratin-associated protein domain (IPR002494). By using ScanProsite, we were able to find more than half a million of such motifs in the UniProtKB database [57]. Manual inspection of the viral LCR fragment shows the presence of a similar C–C–S–C motif. This fragment is also present in more than 500,000 sequences in UniProtKB. Interestingly, all 13 hits to the human proteome are metallothioneins with very similar motifs that are responsible for metal binding [58, 59].

Nsp3 is the largest multi-domain protein encoded by the coronavirus genome. LCR of nsp3 (PPDEDEEEGDCEEEEFE) lies across the borders of two domains identified in coronaviruses: Ub1 (1-112) and acidic domain hypervariable region (HVR) (113–183) [60–62]. This LCR is significantly similar to the Armadillo-type fold (IPR016024), 'a multi-helical fold comprised of two curved layers of alpha helices arranged in a regular right-handed superhelix, where the repeats that make up this structure are arranged about a common axis. These superhelical structures present an extensive solvent-accessible surface that is well suited to binding large substrates such as proteins and nucleic acids [63, 64] https://www.ebi.ac.uk/interpro/entry/InterPro/IPR016024/.

#### Non-recommended epitopes of SARS-CoV-2

In the last section we investigated the lists of epitopes suggested previously [22]; [16, 65]. The authors of those papers provide predictions for 3295 possible candidates for T-cell epitopes and 1519 possible candidates for B-cell epitopes. The epitopes for T or B cells

may be linear or structural (conformational). Linear epitopes consist of linear amino acid (aa) sequence while structural are based on folded protein structure where particular aa comes close to each other in structure. By analysing this data we found that 21 of the predicted T-cell epitopes and 27 (1,7%) of the predicted B-cell epitopes overlap with 5 SARS-CoV-2 LCRs that are significantly similar to human proteins. However, only the S and N proteins from SARS-CoV are known to induce potent and long-lived immune responses [66–71]. This narrows the number of potential candidates to 562 (419 for S protein and 143 for N protein) for T-cell epitopes and to 397 (317 for S protein and 80 for N protein) for B-cell epitopes. Among these, we found out that 11 (2%) of the predicted T-cell epitopes and 19 (5%) of the predicted B-cell epitopes overlap with SARS-CoV-2 LCRs. The lists of B-cell and T-cell overlapping epitopes are presented in Tables 2 and 3 respectively and the overlapping fragments are marked in red colour. We therefore speculate that these regions should not be taken into account while selecting epitopes.

## Discussion

Anti-COVID-19 vaccine development is mainly based on: DNA and RNA technology, peptides, virus-like particle, recombinant protein, viral vector, live attenuated virus and inactivated virus platforms [72]. Although the epitopes for neutralizing SARS-CoV-2 antibody are known, the public information about the specific antigens which were used in vaccine development is not available. Some vaccines are based on S protein or even on whole virion [73]. Based on our findings in SARS-CoV-2 proteins 5 LCRs common for virus and human proteins are presented, clearly indicating that antigens for SARS-CoV-2 vaccine development need to be designed and defined with extreme care. In the case of SARS-CoV-2 and other coronaviruses, the development of effective vaccine is not trivial. First of all, the proper antigen design is critical. This may enable avoidance of such side effects of vaccine as autoimmune disease. The LCR is known as one of the strongest and most immunogenic epitopes and can enhance the immune evasion of pathogen [35]. Additionally, similar LCRs were found in the human proteome. Therefore, LCR used as antigen (1) may generate ineffective antibody (not blocking virus entry into the cell), and (2) may produce the antibody that can serve as the basis for development of autoimmune diseases. Moreover, in the case of coronaviruses, the antigen dependent enhancement (ADE) of virus entry was observed [74]. Therefore, the harmful antibody developed against the not properly designed epitope may potentially cause ADE of SARS-CoV-2. In lentivirus and HIV-1 the LCRs are potentially hypervariable regions and may contribute to the retroviral ability to avoid the immune system [75]. Thus, in conclusion during the design process of the antigen used as the basis for efficient vaccine, the sequences should be carefully investigated for the presence of LCRs which may cause potential harmful effects of the produced vaccine. Based on vaccine development against SARS-CoV and MERS some concerns were recognized including induction of ADE as not neutralizing antibody enhanced virus infectivity. ADE was found in cats vaccinated against a species-specific coronavirus [76]. In case of SARS, the use of whole inactivated virus or S glycoprotein induced hepatitis and lung immunopathology in animal models, while inactivated MERS in vaccination caused pulmonary infiltration in mice [77]. Moreover, it is still unclear whether adaptive T cell responses may also play a

**Table 2** SARS-CoV-2 low complexity regions that overlap with B-cell epitopes

| | Protein | Epitope | LCR | Epitope start | Epitope end | LCR start | LCR end | % of coverage |
|---|---|---|---|---|---|---|---|---|
| * | S | MVTIMLCCMTS | MLCCMTSCCSCLKGCCSCGSCC | 1229 | 1239 | 1233 | 1254 | 64% |
| * | N | NKHIDAYKTFPPTEPKKDKKKKTDEAQPLPQRQKKQPTVTLLPAADM | KTFPPTEPKKDKKKKADE | 354 | 400 | 361 | 378 | 38% |
| * | N | RGGSQASSRSSSRSRNSSRNSTPGSSRGTSPARMAGNGG | GSRGGSQASSRSSSRSRNSSRNSTPGSSRGTSP | 177 | 215 | 175 | 207 | 79% |
| ** | N | QLPQGTTLPKGFYAEGSRGGSQ | GSRGGSQASSRSSSRSRNSSRNSTPGSSRGTSP | 160 | 181 | 175 | 207 | 32% |
| ** | N | PKGFYAEGSRGGSQASSR | GSRGGSQASSRSSSRSRNSSRNSTPGSSRGTSP | 168 | 185 | 175 | 207 | 61% |
| ** | N | SRGGSQASSRSSSRSR | GSRGGSQASSRSSSRSRNSSRNSTPGSSRGTSP | 176 | 191 | 175 | 207 | 100% |
| ** | N | KHIDAYKTFPPTEPKKDKKK | KTFPPTEPKKDKKKKADE | 355 | 375 | 361 | 378 | 67% |
| ** | N | LNKHIDAYKTFPPTEPK | KTFPPTEPKKDKKKKADE | 353 | 369 | 361 | 378 | 43% |
| ** | N | KTFPPTEPKKDKKKK | KTFPPTEPKKDKKKKADE | 361 | 375 | 361 | 378 | 100% |
| ** | N | TFPPTEPK | KTFPPTEPKKDKKKKADE | 362 | 369 | 361 | 378 | 100% |
| *** | nsp3 | YPPDEDEEEG | PPDEDEEEGDCEEEEFE | 107 | 116 | 108 | 124 | 90% |
| *** | nsp3 | PDEDEEEGDCEEEEFEPSTQYEYGTEDDYQGKPLEFGATSAALQPEEEQEEDWLDDDSQQTVGQQDGSEDNQTTTIQTIVEVQPQLEMELTPVVQTIEVNSFSGYLKLT | PPDEDEEEGDCEEEEFE | 109 | 217 | 108 | 124 | 16% |
| *** | nsp3 | PDEDEEEGDCEEEEFEPSTQYEYGTEDDYQGKPLEFGATSAALQPEEEQEEDWLDDDSQQTVGQQDGSEDNQTTTIQTIVEVQPQLEMELTPVVQTIEVNSFSGYLKLT | QPEEEQEEDWLDDDSQ | 109 | 217 | 152 | 167 | 15% |

role in conferring protection against SARS-CoV-2. For SARS-CoV, in human survivors the memory T cells, but not B cells, were found around 6 years after infection [78]. The recent study indicated that in COVID-19 patient the 45 various antibodies against SARS-CoV-2 were found although only 3 exhibited ability to neutralize the

**Table 2** (continued)

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| *** | nsp3 | GDCEEEEFEP | PPDEDEEEGDCEEEEFE | 116 | 125 | 108 | 124 | 90% |
| *** | nsp3 | EEFEPSTQYE | PPDEDEEEGDCEEEEFE | 121 | 130 | 108 | 124 | 40% |
| *** | nsp3 | EEFEPSTQY | PPDEDEEEGDCEEEEFE | 121 | 129 | 108 | 124 | 43% |
| *** | nsp3 | AALQPEEEQE | QPEEEQEEDWLDDDSQ | 148 | 157 | 152 | 167 | 70% |
| *** | nsp3 | EEQEEDWLDD | QPEEEQEEDWLDDDSQ | 163 | 172 | 152 | 167 | 100% |
| *** | S | IAGLIAIVMVTIMLCCMTSCCSCLKGCCSCGSCCKF | MLCCMTSCCSCLKGCCSCGSCC | 1221 | 1256 | 1233 | 1254 | 64% |
| *** | N | TLPKGFYAEGSRGGSQASSRSSSRSRNSSRNSTPGSSRGTSPARMAGNGG | GSRGGSQASSRSSSRSRNSSRNSTPGSSRGTSP | 166 | 215 | 175 | 207 | 66% |
| *** | N | GFYAEGSRGG | GSRGGSQASSRSSSRSRNSSRNSTPGSSRGTSP | 170 | 179 | 175 | 207 | 60% |
| *** | N | GGSQASSRSS | GSRGGSQASSRSSSRSRNSSRNSTPGSSRGTSP | 178 | 187 | 175 | 207 | 100% |
| *** | N | ASSRSSSRSR | GSRGGSQASSRSSSRSRNSSRNSTPGSSRGTSP | 182 | 191 | 175 | 207 | 100% |
| *** | N | RNSSRNSTPG | GSRGGSQASSRSSSRSRNSSRNSTPGSSRGTSP | 191 | 200 | 175 | 207 | 100% |
| *** | N | TPGSSRGTSP | GSRGGSQASSRSSSRSRNSSRNSTPGSSRGTSP | 198 | 207 | 175 | 207 | 100% |
| *** | N | GTSPARMAGN | GSRGGSQASSRSSSRSRNSSRNSTPGSSRGTSP | 204 | 213 | 175 | 207 | 100% |
| *** | N | AYKTFPPTEPKKDKKKKADETQALPQRQKKQQTVTLLPAADLDDFSKQLQQSMSSADS | KTFPPTEPKKDKKKKADE | 358 | 415 | 361 | 378 | 26% |

* —epitopes derived from [22], **—epitopes derived from [16], ***—epitopes derived from[65]

virus [79]. Additionally, antibody against SARS-CoV-2 may cause the cross-reactivity with pulmonary surfactant proteins (shared similarity with 13 out of 24 pentapeptides) and development of SARS-CoV-2-associated lung disease [80]. Furthermore, recent study indicated that antibody against S glycoprotein exhibited ability to cross-react with human tissue proteins including: S100B, transglutaminase 3 and 2 (tTG3, tTG2), myelin basic protein (MBP), nuclear antigen (NA), αmyosin, collagen, claudin 5 + 6 and thyroid peroxidase (TPO) [81].

Our work clearly shows similarity of SARS-CoV-2 protein low complexity sequences to human LCRs. We were able to detect similarity in 3 SARS-CoV-2 proteins to several human protein families. This resemblance can be seen in the nsp3, spike protein (S) as well as in the nucleocapsid protein (N). Previous research shows that both S

**Table 3** SARS-CoV-2 low complexity regions that overlap with T-cell epitopes

| | Protein | Epitope | LCR | Epitope start | Epitope end | LCR start | LCR end | % of coverage |
|---|---|---|---|---|---|---|---|---|
| * | nsp3 | GDCEEEEFEPSTQY | PPDEDEEEGDCEEEEFE | 116 | 129 | 108 | 124 | 64% |
| * | N | AYKTFPPTEPK | KTFPPTEPKKDKKKKADE | 359 | 369 | 361 | 378 | 82% |
| ** | S | CMTSCCSCLK | MLCCMTSCCSCLKGCCSCGSCC | 1236 | 1245 | 1233 | 1254 | 100% |
| ** | S | MTSCCSCLK | MLCCMTSCCSCLKGCCSCGSCC | 1237 | 1245 | 1233 | 1254 | 100% |
| ** | N | LLNKHIDAYKTFPPTEPK | KTFPPTEPKKDKKKKADE | 352 | 369 | 361 | 378 | 50% |
| ** | N | YKTFPPTEPKKDKKKK | KTFPPTEPKKDKKKKADE | 360 | 375 | 361 | 378 | 94% |
| ** | N | SQASSRSSSR | GSRGGSQASSRSSSRSRNSSRN STPGSSRGTSP | 180 | 189 | 175 | 207 | 100% |
| ** | N | AEGSRGGSQA | GSRGGSQASSRSSSRSRNSSRN STPGSSRGTSP | 173 | 182 | 175 | 207 | 80% |
| ** | N | IDAYKTFPPTEPKKD | KTFPPTEPKKDKKKKADE | 357 | 371 | 361 | 378 | 69% |
| ** | N | NKHIDAYKTFPPTEP | KTFPPTEPKKDKKKKADE | 354 | 368 | 361 | 378 | 53% |
| ** | N | KTFPPTEPKK | KTFPPTEPKKDKKKKADE | 361 | 370 | 361 | 378 | 100% |
| ** | N | KTFPPTEPK | KTFPPTEPKKDKKKKADE | 361 | 369 | 361 | 378 | 100% |
| *** | nsp3 | EEFEPSTQY | PPDEDEEEGDCEEEEFE | 121 | 129 | 108 | 124 | 43% |
| *** | nsp3 | YAEGSRGGSQASSRS | GSRGGSQASSRSSSRSRNSSRNS TPGSSRGTSP | 172 | 186 | 175 | 207 | 80% |
| *** | nsp3 | RSRNSSRNS | GSRGGSQASSRSSSRSRNSSRNS TPGSSRGTSP | 185 | 194 | 175 | 207 | 100% |
| *** | nsp3 | SSRSRNSSR | GSRGGSQASSRSSSRSRNSSRNS TPGSSRGTSP | 187 | 195 | 175 | 207 | 100% |
| *** | nsp3 | SSRSSSRSR | GSRGGSQASSRSSSRSRNSSRNS TPGSSRGTSP | 183 | 191 | 175 | 207 | 100% |
| *** | nsp3 | SSRNSTPGS | GSRGGSQASSRSSSRSRNSSRNS TPGSSRGTSP | 193 | 198 | 175 | 207 | 100% |
| *** | nsp3 | QASSRSSSR | GSRGGSQASSRSSSRSRNSSRNS TPGSSRGTSP | 181 | 189 | 175 | 207 | 100% |
| *** | nsp3 | SSRGTSPAR | GSRGGSQASSRSSSRSRNSSRNS TPGSSRGTSP | 201 | 209 | 175 | 207 | 78% |
| *** | nsp3 | KTFPPTEPK | KTFPPTEPKKDKKKKADE | 361 | 369 | 361 | 378 | 100% |

* —epitopes derived from [22], **—epitopes derived from [16], ***—epitopes derived from [65]. Red colour represents exact matches, orange colour represents encompassing matches and violet colour represents overlapping matches

and N proteins are known to induce potent and long-lived immune responses against SARS-CoV.

The nsp3 LCR fragments are part of the hypervariable region (HVR) which is Glu-rich. This region, even if so variable, is always present in all Coronoviridae. It is known to interact with nsp6, nsp8, nsp9 and its own C-terminal part, however no function has been assigned to it to date [61, 82]. The same is true for the human transcription factor Myt1l's glutamic acid-rich region which has an unknown role. Of note is the fact that the enrichment of glutamic acid was found as a feature of the highly immunogenic polypeptides [35]. Since Mytl1l is a transcription factor we may hypothesize that its LCR is somehow linked to the general function of binding nucleic acids. Such parallels may be helpful in understanding SARS-CoV-2 processes.

The surface glycoprotein (S) is of utmost interest to the scientific and medical communities because of its presence on the viral particle surface. The LCR identified in this study is a part of the cysteine-rich motif (CRM) present in the S2 domain, in the most C-terminal end of the protein located in the cytoplasm (endodomain) [83]. This sequence has been shown to be palmitoylated which is a critical step towards incorporation of S to the viral envelope [84–91]. Similarities to keratin-associated proteins and metallothioneins are hard to interpret. There are many possible explanations. One of them is the presence in epithelium. The function of this set of cysteines demands a more detailed study. Buonvino and Melino suggest a hypothetic active role of the coronavirus S protein cytoplasmic domain in protein–protein aggregation for clots formation and cell–cell fusion SARS-CoV-2-S protein-driven [92].

The nucleoprotein/nucleocapsid phosphoprotein (N) packages the viral genome into a helical ribonucleocapsid (RNP) and is crucial during viral self-assembly as shown in experiments with previously known coronaviruses [93–98]. Both regions of interest are located in the SR-rich region of the linkage region (LKR: residues 176–204) and the C-terminal disordered region (residues 370–389) that together with the N-terminal part are involved in RNA binding [99, 100]. Similarity of the N protein to RANB2, an element of the supraspliceosome, seems surprising. However a hypothesis based on results from zebrafish may point at RANB2 as a weapon against infections, as is the case of the fish ZRANB2 [101]. The C-terminal LCR is similar to the human MICAL3 LCR which is multifunctional [48, 49]. Gene Ontology analyses studies appear to indicate an intriguing over-representation of transport functions among human proteins whose LCRs are similar to coronavirus proteins.

It is known that viruses attack major cellular processes like vesicular trafficking, cell cycle, cellular transport, protein degradation and signal transduction to realize their goals [102]. Many host processes are taken over by viral proteins with the use of short linear motifs that are often parts of intrinsically disordered regions (IDRs). For example, the RGD motif mimics the regular cellular machinery for cell attachment via integrin [103]. Many IDRs are composed of low complexity regions. Therefore the hypothesis of the importance of similarities described above are not unfounded. Thorough analysis of SARS-CoV-2 short linear motifs has been recently published by the Gibson's group [104].

The most important outcome of this work is the indication that epitopes cannot be selected based only on factors like phylogenetic conservation or potential epitope

targets. For the safety of patients and procedures, all epitopes that may be similar to human proteome fragments should be discarded from further studies because the cure against SARS-CoV-2 may as well turn against the host.

Due to the fact that several research groups are working on the development of vaccine against SARS-CoV-2 it is very important to highlight the possible weak points which may cause unexpected side effects. The autoimmune diseases rate increased significantly in recent years. Moreover, it correlates with vaccination programmes [105]. Several studies indicated that vaccine components may induce autoimmune disease e.g. vaccine against Lyme disease can cause chronic arthritis and rheumatic heart disease [106]. However the mechanism triggering autoimmune disease after vaccination still remains unclear [107].

We also note a complete lack of LCRs in proteins originating from the nsp9-nsp16 proteins (Fig. 1). Previous studies have shown that LCRs are more often present on protein ends [108], which are hard to define in polyproteins as in the case of pp1ab. The only distinguishing feature of these proteins is their function; most proteins from this group are involved in replication [109]. We speculate that the similarity of viral LCRs to human proteins may not be purely accidental but may be a molecular disguise. We suggest that SARS-CoV-2 may use these regions for specific functions that replace the cellular machinery for its own purposes.

Here we provide the scientific community with tools that allow the comparison of all types of low complexity fragments. These techniques have been shown to be useful previously in order to detect previously unknown similarities (Kubáň et al., 2019; Tørresen et al., 2019) and based on previous results we decided to use these tools to search for similarities among human and SARS-CoV-2 low complexity regions.

LCRs appear to come in 3 flavours. They can consist of homogenous polyX regions (homorepeats), repetitive fragments, or irregular LCRs [110]. Secondly, they usually come in specific combinations of amino acids, e.g. hydrophobic, cysteine-rich (alone or in combination with histidines), and glutamic acid always goes with aspartic acid. Our methods are tailored to detect the different types of such low complexity regions.

The reader of this work should be aware that our results are based on sequence similarity only. We are fully aware that we do not include possible topological similarities of epitopes. These structural resemblances may of course play a role in comparison of even phylogenetically and fold-wise distant protein structures, as shown in allergic cross-reactivity [111, 112]. We therefore cannot exclude that other similarities exist between SARS-CoV-2 and human proteins that are not identified here.

## Conclusions

Finding of five low complexity regions (LCRs) in three SARS-CoV-2 encoded proteins (nsp3, S and N) that are highly similar to regions from human proteome poses a serious threat to the vaccine or drug design. Similarity of SARS-CoV-2 LCRs to human proteins may have implications on the ability of the virus to counteract immune defense. The vaccine targeting LCRs may potentially be ineffective or alternatively lead to autoimmune diseases development.

## Methods

### SARS-CoV-2 protein sequences

All full-length protein sequences of the SARS-CoV-2 proteome were retrieved on 28 April 2020 from the ViralZone web portal (https://viralzone.expasy.org/8996) which provides pre-release access to the SARS Coronavirus 2 protein sequences in UniProt. The UniProtIDs of the SARS-CoV-2 proteins are P0DTC1 replicase polyprotein 1a (pp1a), P0DTD1 Replicase polyprotein 1ab (pp1ab), P0DTC2 Spike glycoprotein (S), P0DTC3 ORF3a protein (NS3a), P0DTC4 Envelope small membrane protein (E), P0DTC5 Membrane protein (M), P0DTC6 ORF6 protein, P0DTC7 ORF7a protein, P0DTD8 ORF7b protein, P0DTC8 ORF8 protein, P0DTC9 Nucleoprotein (N), P0DTD2 ORF9b protein, P0DTD3 ORF14 protein and A0A663DJA2 hypothetical ORF10 protein. Based on the information derived from UniProt replicase polyprotein 1a and replicase polyprotein 1ab were then divided into proteinases responsible for the cleavages of the polyproteins, that is: nsp1, nsp2, nsp3, nsp4, 3C-like proteinase, nsp6, nsp7, nsp8, nsp9, nsp10, nsp11, RNA-directed RNA polymerase, helicase, proofreading exoribonuclease and 2-O methyltransferase.

### Identification of LCRs

To identify low complexity fragments in SARS-CoV-2 proteins we used the PlatoLoCo metaserver [31] which provides a web interface to a set of state-of-the-art methods that allow detection of LCRs, compositionally biased protein fragments, and short tandem repeats. Using all these methods we were only able to detect low complexity protein fragments using the SEG algorithm using the default parameters (W = 12, K 1 = 2.2, K 2 = 2.5). To identify low complexity protein fragments in the human genome we downloaded human proteome from the Uniprot database (UP000005640) and analysed it using SEG with the same set of parameters.

Having LCRs identified based on the proteins derived from the reference SARS-CoV-2 genome we have also analysed the number of mutations already discovered for each AA position in the regions of our interest in order to check if those regions are characterized by a high mutation rate. To perform such analysis we downloaded mutation data from the COVIDep [113], 114. As for the date of data accession (January 8th, 2021) COVIDep database included 232,735 analysed SARS-CoV-2 sequences and CoV-GLUE database included 242,865 SARS-CoV-2 sequences. Based on the obtained information we computed the percentage of the sequences with mutations for each residue of detected LCRs. We notice that for most of the residues the number of sequences with mutations is below 1%. The only exceptions are residues 194, 199, 203, 204 and 365, 376, 377 from two LCR fragments from nucleoprotein (N) where mutation percentages are 5,8%, 3%, 33,6%, 33,6%, and 1.7%, 2,6%, 1,4% in case of the COVIDep database, and 5,2%, 1,7%, 36,9%, 36,8%, and 1,2%, 1,7%, 1% in case of the CoV-GLUE database, respectively. However, due to way greater resistance of LCRs to mutations this kind of change does not seem to be crucial for their functions [115, 116].

The detailed information of the percentage of sequences with mutations from the COVIDep and the CoV-GLUE databases for each residue of detected SARS-CoV-2 LCRs is provided in the Supplementary Material S9.

### Searching for human protein fragments similar to SARS-CoV-2 LCRs

To detect human sequences that are similar to SARS-CoV-2 LCRs we used our three methods: GBSC, MotifLCR and LCR-BLAST (e-value threshold 0.001). The list of human LCRs that are similar to virus LCRs obtained with GBSC, MotifLCR and LCR-BLAST are presented in a Additional files 1, 2, 3: tables S1, S2 and S3, respectively. For GBSC we used default parameters (score threshold 3, distance threshold 7). The method uses whole protein sequences as an input and then identifies repetitive regions that consistof homopolymers or STRs. Then, similar protein fragments are clustered together and each cluster represents particular repetitive patterns. As a result we obtained two clusters that included both virus and human sequences. MotifLCR and LCR-BLAST require low complexity fragments as input. In our case these sequences were obtained using the SEG tool as described above. In the first step MotifLCR removes unique 2-mers in each sequence in order to create artificial sequences, then it searches for repeats in these new sequences and in the last step it creates clusters with native sequences that contain tandem repeats in artificial sequences. Repeat is defined as at least 3 times the occurrence of a specific amino acid pattern.

MotifLCR results consisted of 20 clusters that represented different repetitive motifs. However, the repetitive motifs in the obtained clusters were not specific. Therefore to further narrow down the sequences we used the results of MotifLCR as a subject database for LCR-BLAST and a list of viral LCRs as a query set. Finally, as a third tool we used LCR-BLAST with the viral LCRs as a query set and all human proteome LCRs as a subject database. As a result both MotifLCR and LCR-BLAST returned five clusters each with human LCRs sequences similar to SARS-CoV-2 LCRs.

### Comparing SARS-CoV-2 LCRs to epitopes

Having selected virus fragments that are similar to human sequences we then investigated the lists of T-cell and B-cell epitopes suggested by [22]; [16, 61] in their works. The authors of the first work provide beginning and end amino acid coordinates for each epitope as well as a name of the virus protein and based on this information we were able to identify epitope regions that overlap with SARS-CoV-2 LCRs. In case of the list of epitopes provided by [16] and [61] we used WU-BLAST (http://blast.wustl.edu) with no gaps and parameters optimized for short sequences to find epitopes that align with 100% identity to SARS-CoV-2 LCRs and threshold of minimum length of aligned fragment of 4AA (Additional files 9, 10).

### Gene Ontology enrichment

Gene Ontology enrichment functional analyses were performed on 12 clusters that included sequences similar to SARS-CoV-2 LCRs. Since some proteins may contain more than one LCR, and each of these LCRs may appear in the cluster, in order to avoid redundancy, enrichment analyses have been performed on lists of unique protein sequences. Reference sets for statistical analyses were created depending on the method used to generate clusters. In the case of GBSC to create a reference set we used all 11,361 unique human proteins that composed all other clusters found by the method. In the case of MotfLCR as well as in the case of LCR-BLAST we used the same protein sets that

were used to create bastp search databases and the sizes of the reference sets were 33,880 and 45,068 proteins respectively. To annotate human proteins with their corresponding GO terms from Biological Process, Molecular Function and Cellular Component namespaces we used BiomaRt R package [102]. Statistical analysis was performed with topGO R package [103] and to assess overrepresentation of GO term annotations in obtained clusters we applied hypergeometric test with false discovery Benjamin-Hochberg multiple testing correction with adjusted p-value cutoff 5%.

## Supplementary Information

The online version contains supplementary material available at https://doi.org/10.1186/s12859-021-04017-7.

**Additional file 1. Table S1**: List of human LCRs similar to SARS-CoV-2 LCRs obtained from GBSC.

**Additional file 2. Table S2**: List of human LCRs similar to SARS-CoV-2 LCRs obtained with MotifLCR—> LCR-BLAST.

**Additional file 3. Table S3**: List of human LCRs similar to SARS-CoV-2 LCRs obtained with LCR-BLAST.

**Additional file 4. Table S4**: GO terms enriched for human proteins from all clusters obtained for the SARS-Cov-2 protein nsp3 LCR fragment 108–124.

**Additional file 5. Table S5**: GO terms enriched for human proteins from all clusters obtained for the SARS-Cov-2 protein nsp3 LCR fragment 152–167.

**Additional file 6. Table S6**: GO terms enriched for human proteins from all clusters obtained for the SARS-Cov-2 spike glycoprotein (S) S LCR fragment 1233–1254.

**Additional file 7. Table S7**: GO terms enriched for human proteins from all clusters obtained for the SARS-Cov-2 nucleocapsid protein (N) LCR fragment 175–207.

**Additional file 8. Table S8**: GO terms enriched for human proteins from all clusters obtained for the SARS-Cov-2 protein N LCR fragment 361–378.

**Additional file 9. Table S9**:Percentage of the sequences with mutations from the COVIDep and the Cov-GLUE databases for each residue of detected SARS-CoV-2 LCRs.

**Additional file 10. Table S10**: Links to data and code generated or analyzed during this study.

**Acknowledgements**
We thank Anna Muszewska, Eliana Kaminska, Miguel Andrade, Matthew Merski and Krzysztof Pawłowski for reading and commenting on the manuscript. We thank David Wotton for helpful discussions on Myt1l. We are grateful to all the brave scientists and technicians for the acquisition of viral RNA and their analyses.

**Authors' contributions**
A.G. performed analyses and wrote the paper; P.J and J.Z.-L. performed analyses, E.S. and T.J.S. analyzed the data and wrote the paper, and M.G. conceived the project, wrote the paper, performed analyses and overall direction for the study. All authors read and approved the final manuscript.

**Funding**
Not applicable.

**Availability of data and materials**
There are no additional materials created associated with this study. All data presented and analyzed in the present study was retrieved from the UniProt as described above. The published article includes all data and code generated or analyzed during this study, and they are summarized in the accompanying tables, figures and Supplemental Material S10.

**Ethics approval and consent to participate**
No human/patients or human data involved in the study.

**Consent for publication**
Not applicable.

**Competing interests**
We declare no conflict of interest.

**Author details**
[1] Department of Computer Networks and Systems, Silesian University of Technology, Gliwice, Poland. [2] Institute of Biochemistry and Biophysics, Polish Academy of Sciences, Warsaw, Poland. [3] Department of Molecular and Translational Oncology, Maria Sklodowska-Curie National Research Institute of Oncology, Warsaw, Poland.

Received: 6 November 2020 Accepted: 1 February 2021
Published online: 08 April 2021

**References**

1. Filice GA. SARS, Pneumothorax, and Our Response to Epidemics. Chest. 2004;125:1982–4.
2. Hui DS, Rossi GA, Johnston SL. SARS, MERS and other Viral Lung Infections: ERS Monograph 72. European Respiratory Society; 2016.
3. Li W, Moore MJ, Vasilieva N, Sui J, Wong SK, Berne MA, et al. Angiotensin-converting enzyme 2 is a functional receptor for the SARS coronavirus. Nature. 2003;426:450–4.
4. Xiao X, Chakraborti S, Dimitrov AS, Gramatikoff K, Dimitrov DS. The SARS-CoV S glycoprotein: expression and functional characterization. Biochem Biophys Res Commun. 2003;312:1159–64.
5. Prete M, Favoino E, Catacchio G, Racanelli V, Perosa F. SARS-CoV-2 Inflammatory Syndrome. Clinical Features and Rationale for Immunological Treatment. Int J Mol Sci. 2020;21.
6. Pedersen SF, Ho Y-C. SARS-CoV-2: a storm is raging. J Clin Invest. 2020;130:2202–5.
7. Ye Q, Wang B, Mao J. The pathogenesis and treatment of the `Cytokine Storm' in COVID-19. J Infect. 2020;80:607–13.
8. Soy M, Keser G, Atagündüz P, Tabak F, Atagündüz I, Kayhan S. Cytokine storm in COVID-19: pathogenesis and overview of anti-inflammatory agents used in treatment. Clin Rheumatol. 2020;39:2085–94.
9. Channappanavar R, Perlman S. Pathogenic human coronavirus infections: causes and consequences of cytokine storm and immunopathology. Semin Immunopathol. 2017;39:529–39.
10. Jiang F, Deng L, Zhang L, Cai Y, Cheung CW, Xia Z. Review of the Clinical Characteristics of Coronavirus Disease 2019 (COVID-19). J Gen Intern Med. 2020;35:1545–9.
11. He F, Deng Y, Li W. Coronavirus Disease 2019 (COVID-19): What we know? J Med Virol. 2020;92:719–25.
12. Tang S, Brady M, Mildenhall J, Rolfe U, Bowles A, Morgan K. The New Coronavirus Disease (COVID-19): What Do We Know So Far? https://doi.org/10.20944/preprints202004.0543.v1.
13. Wang L, Wang Y, Ye D, Liu Q. Review of the 2019 novel coronavirus (SARS-CoV-2) based on current evidence. Int J Antimicrob Agents. 2020;55:105948.
14. Andersen KG, Rambaut A, Lipkin WI, Holmes EC, Garry RF. The proximal origin of SARS-CoV-2. Nat Med. 2020;26:450–2.
15. Wen F, Yu H, Guo J, Li Y, Luo K, Huang S. Identification of the hyper-variable genomic hotspot for the novel coronavirus SARS-CoV-2. J Infect. 2020;80:671–93.
16. Ahmed SF, Quadeer AA, McKay MR. Preliminary identification of potential vaccine targets For the COVID-19 coronavirus (SARS-CoV-2) based on SARS-CoV immunological studies. Viruses. 2020;12:254.
17. Saha R, Burra V L. In silico approach for designing of a multi-epitope based vaccine against novel Coronavirus (SARS-COV-2). https://www.biorxiv.org/content/https://doi.org/10.1101/2020.03.31.017459v1. 2020.
18. Lv L, Li G, Chen J, Liang X, Li Y. Comparative genomic analyses reveal a specific mutation pattern between human coronavirus SARS-CoV-2 and Bat-CoV RaTG13. Front Microbiol. 2020;11:3013.
19. Rehman S ur, Shafique L, Ihsan A, Liu Q. Evolutionary trajectory for the emergence of novel coronavirus SARS-CoV-2. Pathogens. 2020;9:240.
20. Shanker A. The possible origins of the novel coronavirus SARS-CoV-2. OSFPreprints. 2020;https://osf.io/a83r4/.
21. Walls AC, Park Y-J, Tortorici MA, Wall A, McGuire AT, Veesler D. Structure, function, and antigenicity of the SARS-CoV-2 spike glycoprotein. Cell. 2020;181:281–92.
22. Grifoni A, et al. A sequence homology and bioinformatic approach can predict candidate targets for immune responses to SARS-CoV-2. Cell Host Microbe. 2020;27(671–80):e2.
23. Vita R, Mahajan S, Overton JA, Dhanda SK, Martini S, Cantrell JR, et al. The immune epitope database (IEDB): 2018 update. Nucleic Acids Res. 2019;47:D339–43.
24. Zhang T, Wu Q, Zhang Z. Probable pangolin origin of SARS-CoV-2 associated with the COVID-19 outbreak. Curr Biol. 2020;30(1346–51):e2.
25. Rangan R, Zheludev IN, Das R. RNA genome conservation and secondary structure in SARS-CoV-2 and SARS-related viruses. bioRxiv 2020;bioRxiv 2020.03.27.012906.
26. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. Basic local alignment search tool. J Mol Biol. 1990;215:403–10.
27. Eddy SR, Mitchison G, Durbin R. Maximum discrimination hidden Markov models of sequence consensus. J Comput Biol. 1995;2:9–23.
28. Tørresen OK, Star B, Mier P, Andrade-Navarro MA, Bateman A, Jarnot P, et al. Tandem repeats lead to sequence assembly errors and impose multi-level challenges for genome and protein databases. Nucleic Acids Res. 2019;47:10994–1006.
29. Kubáň V, Srb P, Štégnerová H, Padrta P, Zachrdla M, Jaseňáková Z, et al. Quantitative conformational analysis of functionally important electrostatic interactions in the intrinsically disordered region of delta subunit of bacterial RNA polymerase. J Am Chem Soc. 2019;141:16817–28.
30. Ziemska-Legiecka J. MotifLCR: motif-based method for clustering low complexity regions (master thesis). 2019. https://apd.uw.edu.pl/diplomas/178134/?_s=1. Accessed 4 May 2020.
31. Jarnot P, Ziemska-Legiecka J, Dobson L, Merski M, Mier P, Andrade-Navarro MA, et al. PlaToLoCo: the first web meta-server for visualization and annotation of low complexity regions in proteins. Nucleic Acids Res. 2020;48:W77–84.
32. Jarnot P, Ziemska-Legięcka J, Grynberg M, Gruca A. LCR-BLAST—A New Modification of BLAST to Search for Similar Low Complexity Regions in Protein Sequences. In: Man-Machine Interactions 6. Springer International Publishing; 2020. p. 169–80.
33. Anand P, Puranik A, Aravamudan M, Venkatakrishnan AJ, Soundararajan V. SARS-CoV-2 strategically mimics proteolytic activation of human ENaC. Elife. 2020;9:e58603.
34. Kebede AM, Tadesse FG, Feleke AD, Golassa L, Gadisa E. Effect of low complexity regions within the PvMSP3α block II on the tertiary structure of the protein and implications to immune escape mechanisms. BMC Struct Biol. 2019;19:6.
35. Hou N, Jiang N, Ma Y, Zou Y, Piao X, Liu S, et al. Low-Complexity Repetitive Epitopes of Plasmodium falciparum Are Decoys for Humoural Immune Responses. Front Immunol. 2020;11:610.

36. Westall FC. Molecular mimicry revisited: gut bacteria and multiple sclerosis. J Clin Microbiol. 2006;44:2099–104.
37. Halpert G, Shoenfeld Y. SARS-CoV-2, the autoimmune virus. Autoimmun Rev. 2020;19:102695.
38. Kreye J, Reincke SM, Prüss H. Do cross-reactive antibodies cause neuropathology in COVID-19? Nat Rev Immunol. 2020;20:645–6.
39. Matsushita F, Kameyama T, Kadokawa Y, Marunouchi T. Spatiotemporal expression pattern of Myt/NZF family zinc finger transcription factors during mouse nervous system development. Dev Dyn. 2014;243:588–600.
40. Mall M, Kareta MS, Chanda S, Ahlenius H, Perotti N, Zhou B, et al. Myt1l safeguards neuronal identity by actively repressing many non-neuronal fates. Nature. 2017;544:245–9.
41. Manukyan A, Kowalczyk I, Melhuish TA, Lemiesz A, Wotton D. Analysis of transcriptional activity by the Myt1 and Myt1l transcription factors. J Cell Biochem. 2018;119:4644–55.
42. Rogers MA, Langbein L, Praetzel-Wunder S, Winter H, Schweizer J. Human hair keratin-associated proteins (KAPs). Int Rev Cytol. 2006;251:209–63.
43. Sperling J, Sperling R. Structural studies of the endogenous spliceosome - The supraspliceosome. Methods. 2017;125:70–83.
44. Mangs AH, Morris BJ. ZRANB2: structural and functional insights into a novel splicing protein. Int J Biochem Cell Biol. 2008;40:2353–7.
45. Hiki K, D'Andrea RJ, Furze J, Crawford J, Woollatt E, Sutherland GR, et al. Cloning, characterization, and chromosomal location of a novel human K+-Cl- cotransporter. J Biol Chem. 1999;274:10661–7.
46. Howard HC, Mount DB, Rochefort D, Byun N, Dupré N, Lu J, et al. The K-Cl cotransporter KCC3 is mutant in a severe peripheral neuropathy associated with agenesis of the corpus callosum. Nat Genet. 2002;32:384–92.
47. Alto LT, Terman JR. MICALs. Curr Biol. 2018;28:R538–41.
48. Grigoriev I, Yu KL, Martinez-Sanchez E, Serra-Marques A, Smal I, Meijering E, et al. Rab6, Rab8, and MICAL3 cooperate in controlling docking and fusion of exocytotic carriers. Curr Biol. 2011;21:967–74.
49. Liu Q, Liu F, Yu KL, Tas R, Grigoriev I, Remmelzwaal S, et al. MICAL3 flavoprotein monooxygenase forms a complex with centralspindlin and regulates cytokinesis. J Biol Chem. 2016;291:20617–29.
50. Frémont S, Romet-Lemonne G, Houdusse A, Echard A. Emerging roles of MICAL family proteins-from actin oxidation to membrane trafficking during cytokinesis. J Cell Sci. 2017;130:1509–17.
51. Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, et al. Gene ontology: tool for the unification of biology. Gene Ontol Consortium Nat Genet. 2000;25:25–9.
52. UniProt Consortium T, The UniProt Consortium. UniProt: the universal protein knowledgebase. Nucleic Acids Research. 2018;46:2699–2699.
53. Sigrist CJA, de Castro E, Cerutti L, Cuche BA, Hulo N, Bridge A, et al. New and continuing developments at PROSITE. Nucleic Acids Res. 2012;41:D344–7.
54. Lu S, Wang J, Chitsaz F, Derbyshire MK, Geer RC, Gonzales NR, et al. CDD/SPARCLE: the conserved domain database in 2020. Nucleic Acids Res. 2020;48:D265–8.
55. Mitchell AL, Attwood TK, Babbitt PC, Blum M, Bork P, Bridge A, et al. InterPro in 2019: improving coverage, classification and access to protein sequence annotations. Nucleic Acids Res. 2019;47:D351–60.
56. Gouw M, Michael S, Sámano-Sánchez H, Kumar M, Zeke A, Lang B, et al. The eukaryotic linear motif resource - 2018 update. Nucleic Acids Res. 2018;46:D428–34.
57. Castro E de, de Castro E, Sigrist CJA, Gattiker A, Bulliard V, Langendijk-Genevaux PS, et al. ScanProsite: detection of PROSITE signature matches and ProRule-associated functional and structural residues in proteins. Nucleic Acids Research. 2006;34 Web Server:W362–5.
58. Krężel A, Maret W. The Functions of Metamorphic Metallothioneins in Zinc and Copper Metabolism. Int J Mol Sci. 2017;18:1237.
59. Sutherland DEK, Stillman MJ. The "magic numbers" of metallothionein. Metallomics. 2011;3:444–63.
60. Neuman BW. Bioinformatics and functional analyses of coronavirus nonstructural proteins involved in the formation of replicative organelles. Antiviral Res. 2016;135:97–107.
61. Lei J, Kusov Y, Hilgenfeld R. Nsp3 of coronaviruses: Structures and functions of a large multi-domain protein. Antiviral Res. 2018;149:58–74.
62. Osipiuk J, Azizi S-A, Dvorkin S, Endres M, Jedrzejczak R, Jones KA, et al. Structure of papain-like protease from SARS-CoV-2 and its complexes with non-covalent inhibitors. bioRxiv. 2020;bioRxiv 2020.08.06.240192.
63. Groves MR, Barford D. Topological characteristics of helical repeat proteins. Curr Opin Struct Biol. 1999;9:383–9.
64. Andrade MA, Perez-Iratxeta C, Ponting CP. Protein repeats: structures, functions, and evolution. J Struct Biol. 2001;134:117–31.
65. Liang C, Bencurova E, Sarukhanyan E, Neurgaonkar P, Scheller C, Dandekar T. Population-Predicted MHCII-Epitope Presentation of SARS-CoV-2 Spike Protein Correlates to the Case Fatality Rates of COVID-19 in Different Countries. 2020. https://papers.ssrn.com/abstract=3576817. Accessed 15 May 2020.
66. Yang Z-Y, Kong W-P, Huang Y, Roberts A, Murphy BR, Subbarao K, et al. A DNA vaccine induces SARS coronavirus neutralization and protective immunity in mice. Nature. 2004;428:561–4.
67. Deming D, Sheahan T, Heise M, Yount B, Davis N, Sims A, et al. Vaccine efficacy in senescent mice challenged with recombinant SARS-CoV bearing epidemic and zoonotic spike variants. PLoS Med. 2006;3:e525.
68. Graham RL, Becker MM, Eckerle LD, Bolles M, Denison MR, Baric RS. A live, impaired-fidelity coronavirus vaccine protects in an aged, immunocompromised mouse model of lethal disease. Nat Med. 2012;18:1820–6.
69. Wang J, Wen J, Li J, Yin J, Zhu Q, Wang H, et al. Assessment of immunoreactive synthetic peptides from the structural proteins of severe acute respiratory syndrome coronavirus. Clin Chem. 2003;49:1989–96.
70. Liu X, Shi Y, Li P, Li L, Yi Y, Ma Q, et al. Profile of antibodies to the nucleocapsid protein of the severe acute respiratory syndrome (SARS)-associated coronavirus in probable SARS patients. Clin Diagn Lab Immunol. 2004;11:227–8.
71. Ng O-W, Chia A, Tan AT, Jadi RS, Leong HN, Bertoletti A, et al. Memory T cell responses targeting the SARS coronavirus persist up to 11 years post-infection. Vaccine. 2016;34:2008–14.
72. Le TT, Andreadakis Z, Kumar A, Román RG, Tollefsen S, Saville M, et al. The COVID-19 vaccine development landscape. Nat Rev Drug Discovery. 2020;19:305–6.

73. Amanat F, Krammer F. SARS-CoV-2 Vaccines: Status Report. Immunity. 2020;52:583–9.
74. Samrat SK, Tharappel AM, Li Z, Li H. Prospect of SARS-CoV-2 spike protein: potential role in vaccine and therapeutic development. Virus Res. 2020;288:198141.
75. María Velasco A, Becerra A, Hernández-Morales R, Delaye L, Jiménez-Corona ME, Ponce-de-Leon S, et al. Low complexity regions (LCRs) contribute to the hypervariability of the HIV-1 gp120 protein. J Theor Biol. 2013;338:80–6.
76. Vennema H, de Groot RJ, Harbour DA, Dalderup M, Gruffydd-Jones T, Horzinek MC, et al. Early death after feline infectious peritonitis virus challenge due to recombinant vaccinia virus immunization. J Virol. 1990;64:1407–9.
77. Padron-Regalado E. Vaccines for SARS-CoV-2: Lessons from Other Coronavirus Strains. Infectious Diseases and Therapy. 2020;9:255–74.
78. Tang F, Quan Y, Xin Z-T, Wrammert J, Ma M-J, Lv H, et al. Lack of peripheral memory B cell responses in recovered patients with severe acute respiratory syndrome: a six-year follow-up study. J Immunol. 2011;186:7264–8.
79. Seydoux E, Homad LJ, MacCamy AJ, Rachael Parks K, Hurlburt NK, Jennewein MF, et al. Analysis of a SARS-CoV-2 infected individual reveals development of potent neutralizing antibodies to distinct epitopes with limited somatic mutation. Immunity. 2020;53(98–105):e5.
80. Kanduc D, Shoenfeld Y. On the molecular determinants of the SARS-CoV-2 attack. Clinical Immunology. 2020;215:108426.
81. Vojdani A, Kharrazian D. Potential antigenic cross-reactivity between SARS-CoV-2 and human tissue with a possible link to an increase in autoimmune diseases. Clin Immunol. 2020;217:108480.
82. Imbert I, Snijder EJ, Dimitrova M, Guillemot J-C, Lécine P, Canard B. The SARS-Coronavirus PLnc domain of nsp3 as a replication/transcription scaffolding protein. Virus Res. 2008;133:136–48.
83. Fung TS, Liu DX. Post-translational modifications of coronavirus proteins: roles and function. Future Virol. 2018;13:405–30.
84. Ujike M, Taguchi F. Incorporation of spike and membrane glycoproteins into coronavirus virions. Viruses. 2015;7:1700–25.
85. Ujike M, Huang C, Shirato K, Matsuyama S, Makino S, Taguchi F. Two palmitylated cysteine residues of the severe acute respiratory syndrome coronavirus spike (S) protein are critical for S incorporation into virus-like particles, but not for M-S co-localization. J Gen Virol. 2012;93(Pt 4):823–8.
86. Shulla A, Gallagher T. Role of spike protein endodomains in regulating coronavirus entry. J Biol Chem. 2009;284:32725–34.
87. Gelhaus S, Thaa B, Eschke K, Veit M, Schwegmann-Weßels C. Palmitoylation of the Alphacoronavirus TGEV spike protein S is essential for incorporation into virus-like particles but dispensable for S-M interaction. Virology. 2014;464–465:397–405.
88. McBride CE, Machamer CE. Palmitoylation of SARS-CoV S protein is necessary for partitioning into detergent-resistant membranes and cell-cell fusion but not interaction with M protein. Virology. 2010;405:139–48.
89. Ye R, Montalto-Morrison C, Masters PS. Genetic analysis of determinants for spike glycoprotein assembly into murine coronavirus virions: distinct roles for charge-rich and cysteine-rich regions of the endodomain. J Virol. 2004;78:9904–17.
90. Petit CM, Chouljenko VN, Iyer A, Colgrove R, Farzan M, Knipe DM, et al. Palmitoylation of the cysteine-rich endodomain of the SARS–coronavirus spike glycoprotein is important for spike-mediated cell fusion. Virology. 2007;360:264–74.
91. Yang J, Lv J, Wang Y, Gao S, Yao Q, Qu D, et al. Replication of murine coronavirus requires multiple cysteines in the endodomain of spike protein. Virology. 2012;427:98–106.
92. Buonvino S, Melino S. New Consensus pattern in Spike CoV-2: potential implications in coagulation process and cell–cell fusion. Cell Death Discovery. 2020;6:134.
93. McBride R, van Zyl M, Fielding BC. The coronavirus nucleocapsid is a multifunctional protein. Viruses. 2014;6:2991–3018.
94. Chang C-K, Hou M-H, Chang C-F, Hsiao C-D, Huang T-H. The SARS coronavirus nucleocapsid protein: forms and functions. Antiviral Res. 2014;103:39–50.
95. Lu S, Ye Q, Singh D, Villa E, Cleveland DW, Corbett KD. The SARS-CoV-2 Nucleocapsid phosphoprotein forms mutually exclusive condensates with RNA and the membrane-associated M protein. bioRxiv. 2020;bioRxiv 2020.07.30.228023.
96. Jack A, Ferro LS, Trnka MJ, Wehri E, Nadgir A, Costa K, et al. SARS CoV-2 nucleocapsid protein forms condensates with viral genomic RNA. bioRxiv. 2020;bioRxiv 2020.09.14.295824.
97. Perdikari TM, Murthy AC, Ryan VH, Watters S, Naik MT, Fawzi NL. SARS-CoV-2 nucleocapsid protein phase-separates with RNA and with human hnRNPs. EMBO J. 2020;39:e106478.
98. Savastano A, de Opakua AI, Rankovic M, Zweckstetter M. Nucleocapsid protein of SARS-CoV-2 phase separates into RNA-rich polymerase-containing condensates. 2020;bioRxiv 2020.06.18.160648.
99. Chang C-K, Hsu Y-L, Chang Y-H, Chao F-A, Wu M-C, Huang Y-S, et al. Multiple nucleic acid binding sites and intrinsic disorder of severe acute respiratory syndrome coronavirus nucleocapsid protein: implications for ribonucleocapsid protein packaging. J Virol. 2009;83:2255–64.
100. Peng T-Y, Lee K-R, Tarn W-Y. Phosphorylation of the arginine/serine dipeptide-rich motif of the severe acute respiratory syndrome coronavirus nucleocapsid protein modulates its multimerization, translation inhibitory activity and cellular localization. FEBS J. 2008;275:4152–63.
101. Wang X, Du X, Li H, Zhang S. Identification of the Zinc Finger Protein ZRANB2 as a Novel Maternal Lipopolysaccharide-binding Protein That Protects Embryos of Zebrafish against Gram-negative Bacterial Infections. J Biol Chem. 2016;291:4019–34.
102. Davey NE, Travé G, Gibson TJ. How viruses hijack cell regulation. Trends Biochem Sci. 2011;36:159–69.
103. Hussein HAM, Walker LR, Abdel-Raouf UM, Desouky SA, Montasser AKM, Akula SM. Beyond RGD: virus interactions with integrins. Adv Virol. 2015;160:2669–81.

104. Mészáros B, Sámano-Sánchez H, Alvarado-Valverde J, Čalyševa J, Martínez-Pérez E, Alves R, et al. Short linear motif candidates in the cell entry system used by SARS-CoV-2 and their potential therapeutic implications. arXiv. 2020. https://arxiv.org/abs/2004.10274. Accessed 21 May 2020.
105. Vadalà M, Poddighe D, Laurino C, Palmieri B. Vaccination and autoimmune diseases: is prevention of adverse health effects on the horizon? EPMA J. 2017;8:295–311.
106. Steere AC, Malawista SE, Snydman DR, Shope RE, Andiman WA, Ross MR, et al. An epidemic of oligoarticular arthritis in children and adults in three Connecticut communities. Arthritis Rheum. 1977;20:7–17.
107. Albert LJ, Inman RD. Molecular mimicry and autoimmunity. N Engl J Med. 1999;341:2068–74.
108. Coletta A, Pinney JW, Solís DYW, Marsh J, Pettifer SR, Attwood TK. Low-complexity regions within protein sequences have position-dependent roles. BMC Syst Biol. 2010;4:43.
109. Subissi L, Imbert I, Ferron F, Collet A, Coutard B, Decroly E, et al. SARS-CoV ORF1b-encoded nonstructural proteins 12–16: replicative enzymes as antiviral targets. Antiviral Res. 2014;101:122–30.
110. Mier P, Paladin L, Tamana S, Petrosian S, Hajdu-Soltész B, Urbanek A, et al. Disentangling the complexity of low complexity proteins. Brief Bioinform. 2019;2:458–72.
111. Platt M, Howell S, Sachdeva R, Dumont C. Allergen cross-reactivity in allergic rhinitis and oral-allergy syndrome: a bioinformatic protein sequence analysis. Int Forum Allergy Rhinol. 2014;4:559–64.
112. Bonds RS, Midoro-Horiuti T, Goldblum R. A structural basis for food allergy: the role of cross-reactivity. Curr Opin Allergy Clin Immunol. 2008;8:82–6.
113. Ahmed SF, Quadeer AA, McKay MR. COVIDep: a web-based platform for real-time reporting of vaccine target recommendations for SARS-CoV-2. Nat Protoc. 2020;15:2141–2.
114. Singer J, Gifford R, Cotten M, Robertson D. CoV-GLUE: A Web Application for Tracking SARS-CoV-2 Genomic Variation. https://doi.org/10.20944/preprints202006.0225.v1.
115. Lenz C, Haerty W, Golding GB. Increased substitution rates surrounding low-complexity regions within primate proteins. Genome Biol Evol. 2014;6:655–65.
116. Radó-Trilla N, Albà M. Dissecting the role of low-complexity regions in the evolution of vertebrate proteins. BMC Evol Biol. 2012;12:155.

## Publisher's Note